DOCUMENT RESUME

ED 088 957 TM 003 530

AUTHOR Williams, John D.; Lindem, Alfred C.
TITLE Regression Computer Programs for Setwise Regression and Three Related Analysis of Variance Techniques.
PUB DATE [74]
NOTE 17p.; Paper presented at the American Educational Research Association Convention (Chicago, Illinois, April 15-19, 1974)

EDRS PRICE MF-$0.75 HC-$1.50
DESCRIPTORS *Analysis of Covariance; *Analysis of Variance; *Computer Programs; *Multiple Regression Analysis; Prediction; Predictor Variables; Statistical Analysis

ABSTRACT

Four computer programs using the general purpose multiple linear regression program have been developed. Setwise regression analysis is a stepwise procedure for sets of variables; there will be as many steps as there are sets. Covaralt allows a solution to the analysis of covariance design with multiple covariates. A third program has three solutions to the two-way disproportionate analysis of variance: (a) the method of fitting constants, (b) the hierarchical model and (c) the unadjusted main effects solution. The fourth program yields three solutions to the two-way analysis of covariance, with or without proportionality, and with multiple covariates. The three solutions are similar to those described for a two-way analysis of variance with disproportionate cell frequencies. (Author)

4.14

REGRESSION COMPUTER PROGRAMS FOR SETWISE REGRESSION AND THREE RELATED ANALYSIS OF VARIANCE TECHNIQUES

John D. Williams and Alfred C. Lindem
The University of North Dakota

A B S T R A C T

Four computer programs using the general purpose multiple linear regression program have been developed. Setwise regression analysis is a stepwise procedure for sets of variables; there will be as many steps as there are sets. COVARMLT allows a solution to the analysis of covariance design with multiple covariates. A third program has three solutions to the two-way disproportionate analysis of variance: (a) the method of fitting constants, (b) the hierarchical model and (c) the unadjusted main effects solution. The fourth program yields three solutions to the two-way analysis of covariance, with or without proportionality, and with multiple covariates. The three solutions are similar to those described for a two-way analysis of variance with disproportionate cell frequencies.

Four different specialized programs have been developed from the utilization of a general purpose multiple linear regression program. The programs that have been developed by these authors are described, together with an indication of the program availability and a description of the statistical technique.

Setwise Regression Analysis

Setwise regression analysis is a technique which was developed (Williams and Lindem, 1971a) to allow a stepwise solution when the interest is in sets of variables rather than in single variables. Thus, the setwise regression procedure bears a strong resemblance to the stepwise regression analysis, and a disadvantage of the stepwise procedure is overcome.

The usual stepwise procedure becomes inappropriate when there are more than two categories being binary coded. A simple example can be made with religious affiliation. Four categories might be used: Catholic, Protestant, Jewish, and other. Three binary predictors can be made with the first three religious affiliations, and the fourth category can be represented as not having membership in the first three categories. If religious affiliation were used in conjunction with other information, the stepwise procedure would not yield a valid indication of the importance of the set of religious variables. The setwise procedure, on the other hand, would allow a direct approach to such a situation.

The setwise procedure drops one set of variables at a time in a stepwise fashion. There will be as many steps as there are sets. The solution is accomplished by an iterative procedure that allows the R^2 (multiple correlation coefficient squared) term to be maximized at each step in a backward stepwise manner. Once a set is discarded, the set is no longer considered at later steps. One set is discarded at each step, until there is only one set remaining.

As a recent issue of VIEWPOINTS has included a complete solution to a setwise problem (Williams, 1973), an example is omitted here. The documentation for the setwise program is given in Williams and Lindem (1971b).

Analysis of Covariance with Multiple Covariates (COVARMLT)

Analysis of covariance programs are typically available, but many of these programs severely limit the number of covariates, usually to one or two covariates. This limitation is wholly unnecessary. The analysis of covariance can be conceptualized as being completed through the use of two linear models, and a multiple linear regression solution follows in a straight-forward manner.

It is helpful to look at the process of the analysis of covariance as it can be generated through the use of linear models. Before the linear models are developed it is useful to set forth a concrete example. Suppose 15 students are split into three groups of five students each and are assigned to three different methods of learning beginning typewriting. Prior to beginning the instructional period, the students are given an intelligence test and a test of manual dexterity. After the conclusion of the experiment a timed typing test is given. Table 1 contains the information for this analysis.

TABLE 1

ANALYSIS OF COVARIANCE WITH TWO COVARIATES

Post-Test	Intelligence Score	Manual Dexterity	Group 1 = 1 0 otherwise	Group 2 = 1 0 otherwise
35	120	38	1	0
27	98	28	1	0
32	102	32	1	0
29	106	22	1	0
27	94	30	1	0
38	123	43	0	1
25	96	31	0	1
36	108	46	0	1
35	115	40	0	1
31	128	35	0	1
27	90	27	0	0
35	110	31	0	0
19	94	25	0	0
17	95	24	0	0
32	116	33	0	0

Table 1 is constructed so that it might be easily transferred to IBM cards for a solution through the use of multiple regression. The group identifiers are binary coded and are found in columns 4 and 5. The group 1 identifier is given by a 1 in column 4, and the group identifier for group 2 is given by a 1 in column 5. A member of group 3 can be identified by having a 0 in both columns 4 and 5. (If there are k groups, then there will be k-1 binary predictors for the group identifiers.)

To accomplish an analysis of covariance by regression it is first necessary to construct a full model. A full model is essentially a model that contains all the information relevant to a data analysis. The full model for the present situation is:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + e_1, \qquad (1)$$

where

Y = the post-test score,

X_1 = the intelligence test score,

X_2 = the manual dexterity score,

X_3 = 1 if the score is from a member of group 1, 0 otherwise,

X_4 = 1 if the score is from a member of group 2, 0 otherwise,

b_0 = the Y-intercept,

$b_1 - b_4$ = the regression coefficients for $X_1 - X_4$, and

e_1 = the error in prediction with the full model.

If this model is solved using a multiple linear regression routine, part of the output will include the multiple correlation coefficient (R). For the present usage, since a full model is being used, the R value found from the use of equation 1 can be labeled R_{FM}.

A <u>restricted model</u> can be developed using only the covariates as predictor variables:

$$Y = b_0 + b_1X_1 + b_2X_2 + e_2, \qquad (2)$$

where

Y = the post-test score,

X_1 = the intelligence test score,

X_2 = the manual dexterity score,

b_0 = the Y-intercept (this b_0 value will, in general, be different than the b_0 value from equation 1),

$b_1 - b_2$ = the regression coefficients for X_1 and X_2 (these regression coefficients will, in general, be different from the b_1 and b_2 values in equation 1), and

e_2 = the error in prediction with the restricted model.

The restricted model also yields an R value, and it can be labeled R_{RM}.

The F test for the analysis of covariance is given by:

$$F = \frac{(R^2_{FM} - R^2_{RM})/(k-1)}{(1 - R^2_{FM})/(N-C-k)} , \qquad (3)$$

where

k is the number of groups,

N is the number of subjects, and

C is the number of covariates.

Using the full model, an R_{FM} value of .88021 is found. Then, $R^2_{FM} = .77478$. For the restricted model, $R_{RM} = .83961$, so that $R^2_{RM} = .70495$.

Using equation 3,

$$F = \frac{(.77478 - .70495)/2}{(1 - .77478)/(15-3-2)} = 1.55.$$

This F value can be interpreted in the usual way with degrees of freedom equal to 2 and 10.

Finding the Adjusted Means

For two covariates the adjusted mean can be found for each group using equation 4:

$$\bar{Y}_k(adj) = \bar{Y}_k - b_1(\bar{X}_{1k} - \bar{X}_{1T}) - b_2(\bar{X}_{2k} - \bar{X}_{2T}), \qquad (4)$$

where

$\bar{Y}_k(adj)$ = the adjusted criterion mean of the k^{th} group,

$\bar{Y}_k$ = the criterion mean of the k^{th} group,

b_1 = the regression coefficient for the first covariate in the full model,

$\bar{X}_{1k}$ = the overall mean on the first covariate,

b_2 = the regression coefficient for the second covariate in the full model,

$\bar{X}_{2k}$ = the mean of the k^{th} group on the second covariate, and

$\bar{X}_{2T}$ = the overall mean of the second covariate.

Additional covariates can be added with no difficulty in an analogous manner. For the present data, $\bar{Y}_1 = 30$, $\bar{Y}_2 = 33$, $\bar{Y}_3 = 26$, $\bar{X}_{11} = 104$, $\bar{X}_{12} = 114$, $\bar{X}_{13} = 101$, $\bar{X}_{1T} = 106.33$, $\bar{X}_{21} = 30$, $\bar{X}_{22} = 39$, $\bar{X}_{23} = 28$, and $\bar{X}_{2T} = 32.33$.

Also, $b_1 = .19514$ and $b_2 = .63027$ (their values are found directly from the printout for the full model).

$$Y_1(adj) = 30 - (.19514)\left[104 - 106.33\right] - (.63027)\left[30 - 32.33\right] = 31.92.$$

$$Y_2(adj) = 33 - (.19514)\left[114 - 106.33\right] - (.63027)\left[39 - 32.33\right] = 27.30.$$

$$Y_3(adj) = 26 - (.19514)\left[101 - 106.33\right] - (.63027)\left[28 - 32.33\right] = 29.77.$$

The process of adjusting the means can be seen as a way to "control" to some extent the difference on the covariates.

Forming a Summary Table

Forming a summary table for the analysis of covariance when using a regression approach is a relatively straight-forward process. The sum of squares within is found directly from the printout from the full model and is 118.32. The adjusted sum of squares total is given by $SS_T(adj) = SS_T(1 - R^2_{RM})$ where R_{RM} is the multiple correlation between Y and the covariates (the restricted model) which, in the present case, is $R_{RM} = .83961$; also $R^2_{RM} = .70495$. With $SS_T = 525.33$, $SS_T(adj) = 525.33\ (1 - .70495) = 155.00$. The adjusted sum of squares among $SS_A(adj)$ can be found as a residual and is $155.00 - 118.32 = 36.68$. The summary table is given in Table 2.

TABLE 2

SUMMARY TABLE FOR THE ANALYSIS OF COVARIANCE WITH TWO COVARIATES

Source of Variation	df	SS	MS	F
Among	2	36.68	18.34	1.55
Within	10	118.32	11.83	
Total	12	155.00		

It should be clear from this presentation that any number of covariates could be employed in an analysis of covariance. Potential researchers should be cautioned against using the "slop bucket" approach to using a large number of covariates simply because it is possible. In addition to being non-scientific, the use of each covariate does entail the loss of one degree of freedom in the adjusted sum of squares within term. A person could use 25 covariates with ease; he should be familiar enough with the data to make a reasonable interpretation of that data after the adjustments, however. A program has been prepared (Williams and Lindem, 1974a) to accommodate up to 20 covariates (which can be redimensioned to include more covariates if necessary); the program prints out summary tables for the analysis of variance for the criterion scores and an analysis of covariance with the multiple covariates and the adjusted means.

Two-Way Fixed Effects Analysis of Variance with Disproportionate Cell Frequencies

The solution to the disproportionate case of the two-way fixed effects analysis of variance is complicated by the existence of more than one solution, the different solutions being dependent upon the assumptions of the researcher. The present program (Williams and Lindem, 1972) allows for the selection of any (or all) of the following least squares solutions: (a) the method of fitting constants, a commonly accepted solution, described in Scheffé (1959) and Anderson and Bancroft (1952), a method that adjusts each main effect for the other main effect; (b) the hierarchical model (Conen, 1968), which allows for one effect to take precedence over the second effect; the first main effect is unadjusted, and the second main effect is adjusted for the first main effect; and (c) the unadjusted main effects method, in which neither main effect is adjusted for the other main effect. In all three methods, the interaction effect is adjusted for the two main effects. The three least squares methods and the previously mentioned approximate solutions are compared by Williams (1972).

As an example of the solutions to the disproportionate two-way situation, consider the following data in Table 3.

TABLE 3

DATA FOR DISPROPORTIONATE TWO-WAY ANALYSIS OF VARIANCE

Effect	B_1	B_2	B_3
A_1	8 6 4	1 1	6 2
A_2	10	7 5 4 4 3	10 9 7 5 4

To solve for any of the three solutions, four linear models are necessary:

Model I: $Y = b_0 + b_1X_1 + e_1$, (5)

Model II: $Y = b_0 + b_2X_2 + b_3X_3 + e_2$, (6)

Model III: $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + e_3$, (7)

Model IV: $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + e_4$, (8) where

Y = the criterion,

X_1 = 1 if the score is from a member of row 1, and 0 otherwise;

X_2 = 1 if the score is from a member of column 1, and 0 otherwise;

X_3 = 1 if the score is from a member of column 2, and 0 otherwise;

$X_4 = X_1 \cdot X_2$;

$X_5 = X_1 \cdot X_3$;

$b_0 - b_5$ = are regression coefficients (The values for b_0, b_1, b_2 and b_3 will, in general, be different for Models I-IV), and

$e_1 - e_4$ = are the errors in prediction with their respective models.

Table 4 contains a formulation for the regression solutions to the two-way fixed effects analysis of variance with disproportionate cell frequencies.

TABLE 4

REGRESSION FORMULATION FOR THE TWO-WAY ANALYSIS OF VARIANCE

Y	X_1	X_2	X_3	X_4	X_5
8	1	1	0	1	0
6	1	1	0	1	0
4	1	1	0	1	0
1	1	0	1	0	1
1	1	0	1	0	1
6	1	0	0	0	0
2	1	0	0	0	0
10	0	1	0	0	0
7	0	0	1	0	0
5	0	0	1	0	0
4	0	0	1	0	0
4	0	0	1	0	0
3	0	0	1	0	0
10	0	0	0	0	0
9	0	0	0	0	0
7	0	0	0	0	0
5	0	0	0	0	0
4	0	0	0	0	0

The values from the regression program that are useful for completing the analysis of variance are: the sum of squares attributable to regression for Models I, II, III, and IV, and the sum of squares for deviation from regression for Model IV. The R^2 values are also included in Table 5. The total sum of squares is of course available from all four models.

The A effect for the method of fitting constants is the difference between the sum of squares for attributable to regression for Model III and Model II: $SS_A = 80.25 - 37.43 = 42.82$.

Essentially, this process amounts to finding that part of the A effect that is independent of the B effect.

The B effect for this method is the difference between the sum of squares for attributable to regression for Model III and Model I:

$SS_B = 80.25 - 20.36 = 59.89$.

TABLE 5

VALUES FOUND FROM THE REGRESSION ANALYSIS

	df	SS	R^2
Model I (A effect) Attributable to Regression	1	20.36	.15427
Model II (B effect) Attributable to Regression	2	37.43	.28355
Model III (Combined A & B effects) Attributable to Regression	3	80.25	.60796
Model IV (Full Model) Attributable to Regression	5	80.80	.61212
Deviation from Regression	12	51.20	
Total Sum of Squares	17	132.00	

Similarly, this second calculation yields that part of the B effect that is independent of the A effect.

And finally, the interaction is found as the difference between Model IV and Model III: $SS_{AB} = 80.80 - 80.25 = .55$. Thus, the effect found in this manner is the AB effect independent of the A and B effects.

The sum of squares for within is equal to the deviation from regression for Model IV. This information for the data in Table 4 can be put into a summary table (Table 6).

TABLE 6

SUMMARY TABLE FOR THE METHOD OF FITTING CONSTANTS

Source of Variation	df	SS	MS	F
A	1	42.82	42.82	10.03**
B	2	59.89	29.95	7.01**
AB	2	.55	.28	.07
Within	12	51.20	4.27	

$^{**}p < .01$

The method of fitting constants is not a partitioning model. That is, if the sum of squares is totaled, it does not equal the total sum of squares of 132.00 (The total is 154.46).

The Hierarchical Model

The hierarchical model (Cohen, 1968) is a method similar to the method of fitting constants. With this approach, a researcher is required to order the variables in relation to their research interest. For example, a researcher may be most interested in the A, or row effect, less interested in the B, or column effect, and may have little interest in the interaction effect. With this approach, each effect is adjusted only for those effects preceding it in the ordering. Thus, the A effect is found directly, the B effect is adjusted for the A effect, and the AB effect is adjusted for the combined A and B effect. Unlike the previous model, this model is additive in the sense that the sum $SS_A + SS_B + SS_W$ is equal to SS_T. The values for SS_A, SS_B, and SS_W can be found from Table 5: $SS_A = 20.36$, the unadjusted A effect: $SS_B = 80.25 - 20.36 = 59.89$, that part of the B independent of A; $SS_{AB} = 80.80 - 80.25 = .55$, as previously, and $SS_W = 51.20$.

These values are placed in a usual summary table (Table 7).

TABLE 7

SUMMARY TABLE FOR THE HIERARCHICAL MODEL

Source of Variation	df	SS	MS	F
A	1	20.36	20.36	4.77*
B	2	59.89	29.95	7.01**
AB	2	.55	.28	.07
Within	12	51.20	4.27	
Total	17	132.00		

*$p < .05$
**$p < .01$

The results from this analysis are identical to the fitting constants method except for the SS_A term. The interpretation would be somewhat different however, because of the decrease in size of the SS_A term. If, on the other hand, the researcher had chosen his order of experimental interest as B, A, AB, then the F values for the A effect and the AB effect would be unchanged from the fitting constants method, but the B effect would be smaller.

The Unadjusted Main Effects Method

A solution similar to the two previous least squares solutions can be called the unadjusted main effects method. Using this approach, both the A and B effects are found directly, with the interaction found in the same manner as the method of fitting constants and the hierarchical model. The error term (mean square within) is of course the same. The values for SS_A, SS_B, SS_{AB}, and SS_W can be found from Table 5: $SS_A = 20.36$, the unadjusted A effect; $SS_B = 37.43$, the unadjusted B effect; $SS_{AB} = 80.80 - 80.25 = .55$, as previously; and $SS_W = 51.20$.

Table 8 contains the unadjusted main effects method analysis.

TABLE 8

SUMMARY TABLE FOR THE UNADJUSTED MAIN EFFECTS METHOD

Source of Variation	df	SS	MS	F
A	1	20.36	20.36	4.77*
B	2	37.43	18.72	4.88*
AB	2	.55	.28	.07
Within	12	51.20	4.27	

*$p < .05$

If the sum of squares is totaled for Table 8, the total is less than 132.00 because of the suppressor relationship between A and B (the total for Table 8 is actually 109.54). The unadjusted main effects method is identical, as a solution, to the one proposed by Jennings (1967). That Jennings' approach and the unadjusted main effects method yield the same results was shown by Halldorson (1969).

Two-Way Analysis of Covariance with Multiple Covariates and Proportionate or Disproportionate Cell Frequencies

The present program (Williams and Lindem, 1974b) is a generalized two-way fixed effects analysis of covariance program that will allow multiple covariates and/or disproportionality of the cell frequencies. Because the program is general, it can be used whether or not there are multiple covariates and whether or not disproportionality of the cell frequencies exists. As was true of the program documented for the two-way fixed effects analysis of variance with disproportionate cell frequencies, three distinct solutions exist for this analysis of covariance situation: (1) the method of fitting constants, a solution that adjusts each main effect for the covariates and the other main effect; (2) the hierarchical model, which allows one main effect to take precedence over the second main effect; the first main effect is adjusted only for the covariates, and the second main effect adjusted for both the first main effect and the covariates; and (3) the unadjusted main effects method, in which the main effects are adjusted only for the covariates. In all three solutions, the interaction effect is adjusted for the covariates and the two main effects. These three solutions are analogous to the previously documented solutions for the fixed effects analysis of variance with disproportionate cell frequencies.

As an illustrative example, suppose the data is cast in a 2 X 3 table with two covariates. Then the following models could be generated:

Model V: $Y = b_0 + b_6X_6 + b_7X_7 + e_5$, (9)

Model VI: $Y = b_0 + b_1X_1 + b_6X_6 + b_7X_7 + e_6$, (10)

Model VII: $Y = b_0 + b_2X_2 + b_3X_3 + b_6X_6 + b_7X_7 + e_7$, (11)

Model VIII: $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_6X_6 + b_7X_7 + e_8$, (12)

Model IX: $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7 + e_9$. (13)

where

Y, X_1, X_2, X_3, X_4, X_5 and $b_0 - b_5$ are defined as previously given in the solution for disproportionate cell frequencies for a two-way analysis of variance,

X_6 = the score on the first covariate for each subject,

X_7 = the score on the second covariate for each subject,

$b_6 - b_7$ = are regression coefficients for X_6 and X_7 respectively, ($b_0 - b_7$ will, in general, be different for Models V-IX), and

$e_5 - e_9$ = the errors in prediction for Models V-IX.

Then, for the fitting constants solution,

SS_A = the SS for attributable to regression for Model VIII - the SS for attributable to regression for Model VII, (14)

SS_B = the SS for attributable to regression for Model VIII - the SS for attributable to regression for Model VI, (15)

SS_{AB} = the SS for attributable to regression for Model IX - the SS for attributable to regression for Model VIII, (16)

and

SS_w = the SS for deviation from regression for Model IX. (17)

For the hierarchical solution with primary interest in the A effect;

SS_A = the SS for attributable to regression for Model VI -
the SS for attributable to regression for Model V, (18)

SS_B = same as equation 15,

SS_{AB} = same as equation 16, and

SS_W = same as equation 17.

For the unadjusted main effects solution:

SS_A = same as equation 18,

SS_B = the SS for attributable to regression for Model VII -
the SS for attributable to regression for Model V, (19)

SS_{AB} = same as equation 16, and

SS_W = same as equation 17.

The fitting constants solution for the analysis of covariance can be seen as analogous to the fitting constants solution for the two-way analysis of variance, except that the covariates are also removed as a source of variation; thus, the A effect in the fitting constants solution is that portion independent of both the B effect and the covariates. In the hierarchical solution, the effect of primary research interest is adjusted for the covariates only; in the unadjusted main effects solution, the main effects are adjusted for the covariates only, and not adjusted for the other main effect. The interaction effect and within term are the same for all three solutions.

The solutions for COVARMLT (the analysis of covariance with multiple covariates) and the two-way analysis of covariance described here do not include a test for the homogeneity of the regression on the covariates. Future revisions of these two programs will include options for running these tests if the user so desires.

REFERENCES

Anderson, R. L., and Bancroft, T. A. Statistical theory in research. New York: McGraw-Hill, 1952.

Cohen, J. Multiple regression as a general data-analytic system. Psychological Bulletin, 1968, 70, 426-443.

Halldorson, M. H. An iterative multiple regression technique for the factoral analysis of variance with unequal and disproportionate cell frequencies. Unpublished doctoral dissertation, The University of Northern Colorado, 1969.

Jennings, E. Fixed effects analysis of variance by regression analysis. Multivariate Behavioral Research, 1967, 2, 95-108.

Scheffé, H. The analysis of variance. New York: Wiley, 1959.

Williams, J. D. Two-way fixed effects analysis of variance with disproportionate cell frequencies. Multivariate Behavioral Research, 1972, 7, 67-83.

Williams, J. D. Applications of setwise regression analysis. Multiple Linear Regression Viewpoints, 1973, 4, No. 2, 1-7.

Williams, J. D., and Lindem, A. C. Setwise regression analysis--A stepwise procedure for sets of variables. Educational and Psychological Measurement, 1971a, 31, 747-748.

Williams, J. D., and Lindem, A. C. Setwise Linear Regression, UND Computer Center Special Report, No. 30, November, 1971b.

Williams, J. D., and Lindem, A. C. A computer program for two-way fixed effects analysis of variance with disproportionate cell frequencies. UND Computer Center Special Report, No. 35, October, 1972.

Williams, J. D., and Lindem, A. C. A computer program for the analysis of covariance with multiple covariates. UND Computer Center Special Report, March, 1974a.

Williams, J. D., and Lindem, A. C. A computer program for the two-way analysis of covariance with multiple covariates and proportionate or disproportionate cell frequencies. UND Computer Center Special Report, March, 1974b.